Content Tagging Enriching the Source

Brian Talarczyk, AMSEC LLC
Glenn Handrahan, SAIC

TMMA Conference

May 2004

Content Tagging I (Nov 02 – Jun 03)

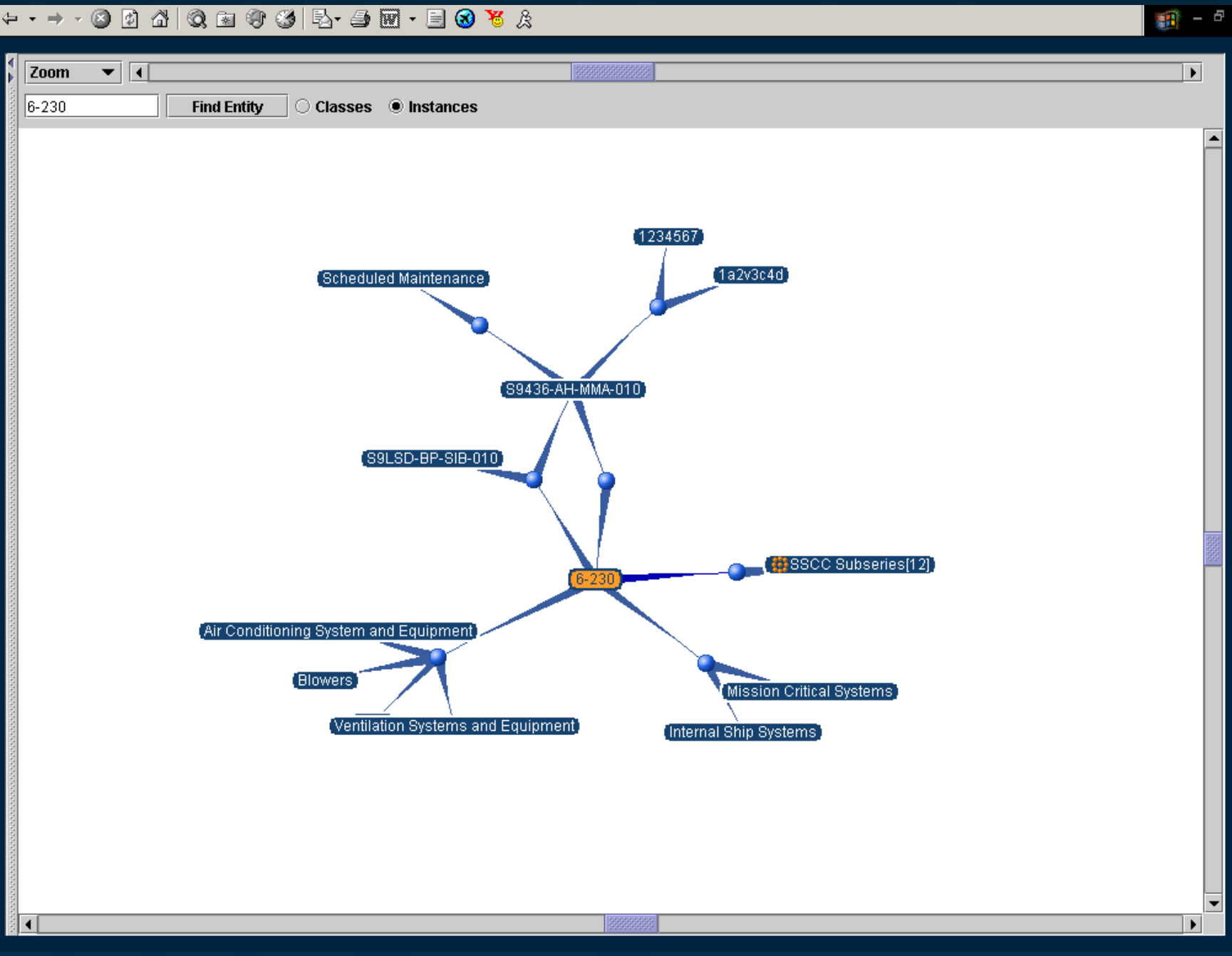
- Increase the intelligence of Navy technical source data
  - Adding content specific metadata
  - Integrating external knowledge resources
  - More relevant search results
- Provided 273 Source Documents for DDG51
  - SIBs
  - Technical Manuals
    - Non-Standard
    - Standard
- Approach
  - LSA
  - Document Analysis

LSA Application #1

- Vendor: H5 Technologies
  - Relies on taxonomies/categorization
  - PhD/SME generated 80,000 categories
  - Content scanned and analyzed to determine its applicability to a category
  - Bar code score given in appropriate category
  - Other documents could be used for a query and its bar code analyzed to determine relevance to source documents
  - Example: SIBs

LSA Application #2

- Vendor: Semagix
  - Algorithms developed to interpret content
  - Employs external resources (Lexis-Nexis), 10 Most Wanted List) to add intelligence to source document
  - SSCC, ESWBS, Slang
  - Visualizer


Zoom
6-230
Find Entity
Classes
Instances
1234567
1a2v3c4d
Scheduled Maintenance
S9436-AH-MMA-010
S9LSD-BP-SIB-010
6-230
SSCC Subseries[12]
Air Conditioning System and Equipment
Blowers
Ventilation Systems and Equipment
Mission Critical Systems
Internal Ship Systems

Document Analysis

- Standard NAVSEA TM – HM&E
  - Eight chapters
    - General Info/Safety Summary
    - Operation
    - Functional Description
    - Preventive Maintenance
    - Troubleshooting
    - Corrective Maintenance
    - IPB
    - Installation
  - SGML Tagged to the NAVSEA C2 DTD

Simple XML vs. HTML Example

- **HTML**

```
<H1>                    </H1>
<p>                 </p>
<p>                 </p>
<p>                 </p>
<p>                 </p>
```

Simple XML vs. HTML Example

- **HTML**

```
<H1>Customer Personal info</H1>
<p>John Smith</p>
<p>211 Main Street</p>
<p>Anytown, VA 22000</p>
<p>703-555-5515</p>
```

- **XML**

```
<customer>
  <name>
    <last>Smith</last>
    <first>John</first>
  </name>
  <address>
    <street>211 Main Street</street>
    <city>Anytown</city>
    <state>VA</state>
    <zip>22000</zip>
  </address>
  <phone>703-555-5515</phone>
</customer>
```

Simple XML vs. NAVSEA C2 SGML Example

- **SGML**

```
<Chapter><title>       </title>
  <Section><title>     </title>
   <Para0><title>      </title>
    <para>    </para>
      <step1>    </step1>
      <step1>    </step1>
      <step1>    </step1>
   </Para0>
</Section>
</Chapter>
```

Simple XML vs. NAVSEA C2 SGML Example

- **SGML (Structure-Based)**

```
<Chapter
    label="6"><title>Corrective
    Maintenance</title>
  <Section><title>Repair</title>
   <Para0><title>Gas Turbine
    Stages</title>
    <para>The Gas Turbine has 14
    stages which need special
    maintenance…</para>
     <step1>Loosen…    </step1>
     <step1>Repair…    </step1>
     <step1>Tighten…    </step1>
   </Para0>
  </Section>
</Chapter>
```

- **XML (Content Tagged)**

```
<Procedure type="Corrective
    Maintenance" action="Repair">
 <title>Gas Turbine Stages</title>
   <Description>The Gas Turbine has
    14 stages which need special
    maintenance…</Description>
     <step1>Loosen…    </step1>
     <step1>Repair…    </step1>
     <step1>Tighten…    </step1>
</Procedure>
```

Content Tagging II (Nov. ’03 to Present)

- LSA Conclusions
  - Too Expensive
  - Too much Customization required
- Convert Legacy Structures to XML Content:
  - Conversion Software Prototype
  - Subject Matter Expert data analysis tools
  - NAVSEA C2 SGML to 24784C XML

Content Model: NAVSEA 24784C XML Spec

- Developed for new Acquisitions (not Legacy)
- Content Specific Tagging:
  - 13 Information Modules (Operation, Maintenance)
  - 54 Information Packages (Procedures, Parts Info)